Fine-grained German Sentiment Analysis on Social Media

Saeedeh Momtazi

Information Systems
Hasso-Plattner-Institut
Potsdam University, Germany
Saeedeh.momtazi@hpi.uni-potsdam.de

Abstract

Expressing opinions and emotions on social media becomes a frequent activity in daily life. People express their opinions about various targets via social media and they are also interested to know about other opinions on the same target. Automatically identifying the sentiment of these texts and also the strength of the opinions is an enormous help for people and organizations who are willing to use this information for their goals. In this paper, we present a rule-based approach for German sentiment analysis. The proposed model provides a fine-grained annotation for German texts, which represents the sentiment strength of the input text using two scores: positive and negative. The scores show that if the text contains any positive or negative opinion as well as the strength of each positive and negative opinions. To this aim, a German opinion dictionary of 1,864 words is prepared and compared with other opinion dictionaries for German. We also introduce a new dataset for German sentiment analysis. The dataset contains 500 short texts from social media about German celebrities and is annotated by three annotators. The results show that the proposed unsupervised model outperforms the supervised machine learning techniques. Moreover, the new dictionary performs better than other German opinion dictionaries.

Keywords: sentiment analysis, sentiment strength, German celebrities, social media, opinion dictionary

1. Introduction

The extraction of sentiment data has become one of the active topics in recent years. This popularity stems from the fact that people like to express their opinions and they are eager to know about other opinions. Web 2.0 technologies provide new resources widely used for such tasks: there are many web services that provide data about opinions on different products; tourism services present opinions about hotels, sightseeings, restaurants, etc; there are also various social media and web pages, such as Twitter, Facebook, blogs, and YouTube, that are used to express opinions about celebrities.

Having such a huge amount of data on the Web, a system that can automatically identify opinions and emotions from text is an enormous help to a user trying to extract such information. Natural language processing applications benefit from being able to distinguish between positive and negative opinions. In addition to this binary classification, users are eager to know about the degree of positivity or negativity of a text. Providing such information requires a fine-grained analysis on opinionated data to find the sentiment strength.

There are many systems developed for classifying texts using machine learning methods. Although learning techniques have been successful in many natural language processing applications, they have rather mixed success for sentiment analysis (Thelwall et al., 2010; Thelwall et al., 2012; Wiegand et al., 2008). In addition, they required a large set of annotated data which is very expensive to provide; while rule-based techniques can achieve a comparable performance without using annotated data. Rule-based models, however, are language dependent and more effort is required to adapt a system to another language. To run a sentiment analyzer for a new language, we need to be familiar with the linguistic behavior of the target language. In addition, we should provide an opinion dictionary to recognize opinionated words within a text. Providing such a dictionary becomes more critical when we aim to consider the sentiment strength of the text. In this paper, we describe a rule-based sentiment analyzer for German which assigns a positive and negative degree to each input text. In order to deal with German data, a dictionary of German opinionated words is provided in which each word is assigned either a positive or a negative degree. In addition, we provided a dataset of 500 German texts in which each sentence has a positive $(0..+3)$ and a negative $(0..-3)$ label. This dataset is used for testing our model and it is publicly available for other applications.[1]

The structure of this paper is as follows: in the next section, we describe our rule-based sentiment analysis model. Section 3 introduces the German opinion dictionary developed for our task. Section 4 introduces the dataset that is provided for this research. The evaluation of our fine-grained German sentiment analyzer is discussed in Section 5. Finally, Section 6 concludes the paper.

2. Rule-based Sentiment Analysis

The main idea of rule-based sentiment analysis is to look for opinionated words in each sentence and classify the input text based on the number of positive and negative words in the text. The list of opinionated words are normally provided as a dictionary, called *opinion dictionary* or *sentiment dictionary*. An opinion dictionary consists of two different lists of words: positive, and negative, which includes all

[1] www.hpi.uni-potsdam.de/fileadmin/hpi/FG_Naumann/bachelorprojekte/BP2011N2/GermanSentimentData.zip

Table 1: Sample positive and negative words from an English opinion dictionary

Positive	Negative
hopefully	petty
cool	bad
happy	fight
love	awful
ecstatic	excruciate

Table 2: Sample words from an English opinion dictionary with fine-grained scores

Positive		Negative	
hopefully	+1	petty	-1
cool	+2	bad	-2
happy	+3	fight	-3
love	+4	awful	-4
ecstatic	+5	excruciate	-5

positive and negative words of an specific language. Table 1 shows a set of sample words of an English opinion dictionary.

For binary polarity classification, the polarity that appears more in the text is assigned to the text. For example, if a sentence has more positive words than the negative ones, it is classified as a positive sentence. For fine-grained classification, the frequency of the opinionated words is also taken into consideration. For example, if a sentence contains 4 positive and 1 negative words, the sentiment strength of the sentence will be $+3$. This value can also be normalized to have a better and comparative representation of sentence sentiments. These approaches, however, are not accurate; because all positive or negative words are considered the same. For example, the sentences "The song was *good*." and "It was an *excellent* song." will be assigned the same sentiment, because both "good" and "excellent" have the same label in the dictionary. To overcome this problem, we need a dictionary with more detailed information about opinionated words such that in addition to the polarity of the words, each word is assigned a sentiment degree. Table 2 presents sample words appeared in an opinion dictionary with fine-grained sentiment scores.

In addition to opinionated words, which are the base part of rule-based sentiment analysis, booster words and negation words play important roles in recognizing the sentiment strength of a sentence.

The sentiment strength in a text is increased or decreased based on the booster words appearing in the sentence. For example, the opinion degree of the following sentences are totally different because of their booster words:

"The song was interesting."

"The song was *very* interesting."

"The song was *somewhat* interesting."

Negation words are also a big concern in sentiment analysis. Occurring a negation word together with an opinionated word flips the polarity of the sentence. For example, the sentence "The song was interesting." which has a positive polarity becomes a negative sentence, in case a negation word like "not" appears close to the opinionated word: "The song was *not* interesting.".

Sentiment analysis becomes even more challenging when we have a combination of both booster and negation words in a single sentence. For example, the sentence "The song was *not very* interesting." is not as negative as the sentence "The song was *not* interesting.". This sample shows that simply flipping the sentiment strength of a sentence is not a good solution for dealing with negation, while the degree should also be changed. For example, although the sentiment strength of the sentence "The song was *very* interesting" is $+3$, the sentence "The song was *not very* interesting" should not be labeled as -3.

Another important challenge in sentiment analysis is classifying texts that have mixed opinions; i.e., sentences that express both positive and negative opinions. As an example, consider the following sentence:

"I found the song very *interesting*, but I think its title was *strange*."

A normal sentiment analyzer will assign either a positive or a negative label to this sentence, while in fact both labels should be assigned to this sentence. Although these kinds of texts are very common in social media, they are not studied deeply. To give an accurate label to such sentences, both positive and negative scoring should be used at the same time; i.e., instead of assigning a single label to each sentence, each sentence will be assigned two labels: positive and negative. In this model, if a text is purely positive, then the negative score will be zero, and if a text is purely negative, then the positive score will be zero. But for texts with mixed sentiment, both positive and negative degrees are non-zero.

To use an opinion dictionary together with the set of booster words and negation words, we utilized the SentiStrength toolkit (Thelwall et al., 2010). The toolkit assigns a sentiment strength to each text based on the opinionated, booster, and negation words in the text. The system is originally developed for English, but it can be adapted to other languages including German. To this aim, a German opinion dictionary as well as a list of German booster and negation words were provided.

In addition to the dictionary, the linguistic behavior of German should also be taken into consideration. For example, in English negation words appear before opinionated words, such as "I do *not* love the song.". In German, however, they can also appear after opinionated words, such as "Ich liebe *nicht* das Lied.". It can also be more complicated when the sentence is formulated as "Ich liebe das Lied *nicht*." in which the negation word does not appear close to the opinionated words. To adapt the toolkit to German, we utilized additional features that cover these phenomena.

3. Opinion Dictionary

As mentioned, an opinion dictionary includes a list of positive words and a list of negative words, which helps sen-

timent analysis systems to label opinionated text. Various opinion dictionaries have been developed for English, including Subjectivity Clues (Wiebe et al., 2005; Wilson et al., 2005), SentiSpin (Takamura et al., 2005), SentiWordNet (Esuli and Sebastiani, 2006), Polarity Enhancement (Waltinger, 2009), and SentiStrength (Thelwall et al., 2010).

There are also a number of opinion dictionaries for German. In 2010, GermanPolarityClues was introduced by Waltinger (Waltinger, 2010) as a German opinion dictionary translated from Subjectivity Clues and SentiSpin dictionaries. This dictionary contains lists of positive and negative words, but it has no opinion degree. At the same time, the SentiWortSchatz dictionary was introduced by Remus (Remus et al., 2010). This dictionary which is the translation of SentiWordNet consists of positive and negative words as well as their sentiment strength. Having a look at the dictionary, we noticed that many opinionated words of German are still missing in this dictionary or their degree of opinion is very low, while the meaning has a stronger sentiment. These shortcoming in available opinion dictionaries for German motivated us to develop a new German dictionary. To this aim, the SentiStrength (Thelwall et al., 2010) dictionary was used, since this dictionary is associated with the SentiStrength toolkit that we use for our sentiment analysis.

To provide the opinion dictionary, the original English dictionary was automatically translated to German by Google translator. The translated words were then manually checked by two native German speakers one of whom is a linguist. While checking the dictionary, four actions were performed:

- keep the word and the score with no changes;
- keep the word and change its score, in case the opinion degree of the German word is not the same as the original English word;
- remove the word, if the German translation has no opinionated sense;
- add new words to the dictionary, in case their English translation has no opinionated sense, while the German words are opinionated.

Overall, the dictionary contains 1,864 words. We performed the same processing for booster and negation words. But the number of words in these two sets are much less than the opinion dictionary. The German booster words contain 54 items and the negation words are 17.

4. Data Annotation

4.1. Dataset

Celebrities are one of the important subjects of opinions on social media. At the end of each World Cup soccer game, a large amount of opinions are expressed on social media about the soccer players contributed to the game; releasing a new album by a singer ends to new opinions that people share with each other about him/her; a new movie which comes out loads new opinions about the actors/actresses who played in that movie. Having such potential opinionated data on social media motivated us to select our dataset from this domain.

Table 3: Data sources used for data annotation

Data Source	# Texts
Facebook	81
Blogs	121
Amazon Comments	138
YouTube Comments	150
Other (Forums, Fan pages)	10

The data collection selected for this study is a set of 500 short texts about German celebrities with a focus on German singers and musicians. On average, there are 1.78 sentences and 20.04 words in each text. The texts are gathered from different social media, including Facebook and blogs. We also used YouTube and Amazon comments on video clips or CDs/DVDs published by celebrities. For YouTube and Amazon comments, we only selected the texts that are about celebrities themselves and not the quality of the clips or CDs/DVDs. Table 3 shows the distribution of our dataset over different social media.

To annotate each text, two scores are required: positive and negative. As mentioned, unlike most of the annotation schema which label texts with one score, our model presents a more detailed and understandable annotation of data, especially for texts with mixed opinions. For each of these degrees, we defined 4 levels: 0 to $+3$ for positive scores, and 0 to -3 for negative scores.

4.2. Annotation Agreement

Since rating opinions is a rather subjective task and it is very difficult to have a reliable annotation for such text data, each text in our dataset is annotated by three people. All annotators are native German speakers.

We calculated inter-annotator agreement on the annotated data with respect to the positive and negative annotations. Figure 1 presents the percentage of agreement between annotators; i.e. the percentage of the data annotated the same by all annotators or two of the annotators. The percentage of agreement is calculated as follows:

$$P(A) = \frac{\text{number of texts agreed}}{\text{total number of texts}} \quad (1)$$

In this figure, the black part shows the percentage of the text that all three annotators agreed on the label. The dark grey part shows the percentage of the text that only two of the annotators agreed on the label. The light grey part is the percentage of the text labeled differently by each annotator.

In annotation of each text, two different kinds of agreements are considered: (1) agreement on the text polarity, and (2) agreement on the sentiment strength. For the former, all positive degrees are considered the same and are analyzed against zero degrees. We made the same assumption for negative annotation. Based on the polarity agree-

Figure 1: Results of the inter-annotator agreement (a: positive annotation, b: negative annotation)

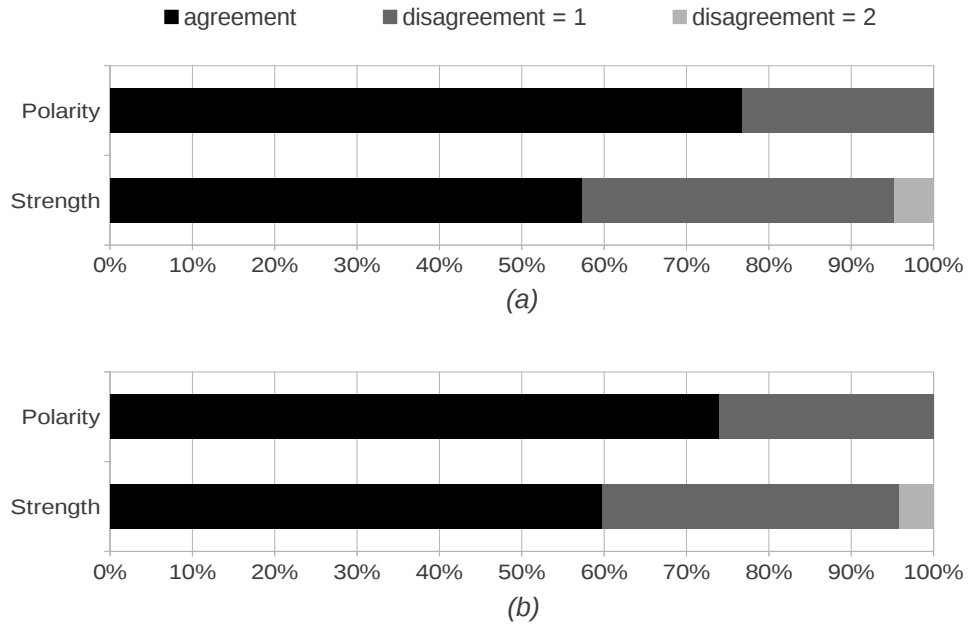

Table 4: Inter-annotator agreement in positive annotation using Kappa

Type	Annotator Pair	P(A)	P(E)	Kappa
Polarity	Annotator 1 & 2	0.862	0.546	0.696
	Annotator 1 & 3	0.796	0.567	0.529
	Annotator 2 & 3	0.878	0.623	0.676
Strength	Annotator 1 & 2	0.742	0.394	0.574
	Annotator 1 & 3	0.662	0.386	0.449
	Annotator 2 & 3	0.696	0.402	0.492

Table 5: Inter-annotator agreement in negative annotation using Kappa

Type	Annotator Pair	P(A)	P(E)	Kappa
Polarity	Annotator 1 & 2	0.858	0.493	0.720
	Annotator 1 & 3	0.780	0.482	0.575
	Annotator 2 & 3	0.842	0.516	0.673
Strength	Annotator 1 & 2	0.760	0.386	0.609
	Annotator 1 & 3	0.684	0.360	0.506
	Annotator 2 & 3	0.710	0.341	0.560

ment, we have an inter-annotator agreement of 76% for positive and 74% for negative annotations; i.e., 76% of positive texts and 74% of negative texts are labeled the same by all three annotators. For the strength agreement, all three levels of scores as well as the zero degree are taken into consideration. Based on this definition, we achieved an agreement of 57% for positive and 60% for negative annotations. These results indicate that rating the strength of an opinion manually is already a challenging task in the annotation process and achieving high performance on automatic fine-grained sentiment analysis is more difficult than coarse-grained analysis.

When calculating the percentage of agreement between two annotators, there is always a portion of data that can be labeled the same randomly; i.e., annotators might agree by chance. To consider this issue and have a normalized measurement of inter-annotator agreement, we need to compute the expected chance agreement and remove it from the percentage of agreement. The expected chance agreement is calculated as follows:

$$P(E) = \sum_{i=1}^{M} (NA_i * NB_i) \quad (2)$$

where M is the number of available labels which is 4 in our case: 0,+1,+2,+3 for positive labeling and 0,-1,-2,-3 for negative labeling. NA_i is the number of texts labeled as i by annotator A, and NB_i is the number of texts labeled as i by annotator B.

Having the percentage of agreement and the expected chance agreement, the $Kappa$ value for inter-annotator agreement is calculated as follows:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (3)$$

where $P(A)$ is the percentage of inter-annotator agreement from Equation 1 and $P(E)$ is the expected chance agreement from Equation 2.

Tables 4 and 5 shows the result of the inter-annotators agreement for positive and negative labeling respectively. As we can see in the tables, the $kappa$ value for polarity labeling is higher than the sentiment strength labeling which

Table 6: Degree of agreement between all annotators using Fleiss' Kappa

	Type	P(A)	P(E)	Kappa
Positive	Polarity	0.845	0.585	0.627
Positive	Strength	0.670	0.399	0.501
Negative	Polarity	0.827	0.504	0.650
Negative	Strength	0.718	0.368	0.554

verifies our observation from Figure 1. Comparing the annotator pairs, the results of positive labeling and negative labeling as well as polarity and strength labeling show that Annotator 1 and 2 have the highest amount of agreement in their opinion, while Annotator 3 has a different idea on labeling opinion texts. Nevertheless, the agreements of Annotator 3 with other annotators are still close to the agreement between Annotator 1 and 2.

Although inter-annotator agreement is only used to assess the agreement between two annotators, the model can also be extended to calculate the degree of agreement between multiple annotators by using *Fleiss' kappa* (Fleiss, 1971). Similar to $Kappa$, this measure also considers the amount of agreement which would be expected by chance and removes this value from the percentage of agreement between all annotators. Table 6 presents the degree of agreement between three annotators using *Fleiss' Kappa*.

5. Evaluation

To evaluate our model, we used the 500 annotated texts described in Section 4. Since each text in the dataset is assigned two individual labels, positive and negative, we considered two separate evaluations for each text. The results of our system are compared to three different supervised machine learning techniques, namely naive Bayes classification, support vector machine, and decision tree. To this aim, we used Weka, a data mining toolkit released by the Machine Learning Group at University of Waikato.[2] In order to train the machine learning models, we used 90% of the dataset as the training set and the rest as the test set. Since our dataset is small for both training and testing, 10-fold cross validation is used. Table 7 presents the results of our experiment.

As can be seen in the table, even though our sentiment analyzer is unsupervised and does not benefit from the annotated data, it significantly outperforms all machine learning methods for negative labeling. Comparing the result of our model with the machine learning methods for positive labeling shows that the performance of the rule-based method is better than naive Bayes classification and decision tree for positive labeling as well, and it is very close to support vector machine.

Overall, the performance of our fine-grained sentiment analyzer is not very high, but considering the results of the inter-annotator agreement presented in Figure 1 and Table 6, we can see that the performance of our system is comparable to the annotator agreement and the test dataset is potentially very difficult and challenging, which makes the task very hard.

[2] http://www.cs.waikato.ac.nz/ml/weka/

Table 7: Performance of the fine-grained German sentiment analysis

Model	Positive	Negative
Rule-based Method	49.0%	58.2%
Naive Bayes	47.6%	46.2%
Support Vector Machine	49.2%	49.0%
Decision Tree	42.2%	44.0%

Table 8: Performance of the coarse-grained German sentiment analysis

Model	Positive	Negative
Rule-based Method	69.6%	71.0%
Naive Bayes	64.0%	62.0%
Support Vector Machine	69.2%	61.4%
Decision Tree	60.0%	54.6%

To compare the difficulty of a fine-grained sentiment analyzer with a typical sentiment analyzer that only recognizes the polarity of sentences, we did a coarse-grained sentiment analysis on the dataset in which all positive and negative labels on the dataset and our system are replaced by $+1$ and -1 respectively; i.e., the system cannot distinguish between texts with different degrees of positivity. The results of this experiment are presented in Table 8.

The results show the superiority of our rule-based model to other approaches even for coarse-grained sentiment analysis. Comparing these results with the results of Table 7 shows that fine-grained sentiment analysis is more difficult than a normal binary sentiment classification, and it is more difficult to achieve good performance on this level of annotation. Although the overall performance of coarse-grained annotation is not very high, our result is still reasonable considering the inter-annotator agreement.

Comparing the performance of our model on positive and negative labeling for both fine-grained and coarse-grained sentiment analyses shows that the accuracy of negative labeling for both fine-grained and coarse-grained approaches is better than the positive labeling. We have the same observation when calculating inter-annotator agreement; i.e. the agreement on negative labeling is higher than the positive labeling (see Tables 4 - 6). This observation shows that the negative opinions that are normally expressed on social media are more transparent than the positive opinions. As a result, it is easier for annotators to label negative texts. The sentiment analysis systems also have a higher accuracy in detecting negative texts.

In addition, we evaluated the German opinion dictionary provided for this study against the other German opinion dictionary. As mentioned, to the best knowledge of the author, there are two dictionaries publicly available

Table 9: Comparing the performance of the GermanSentiStrength and SentiWS opinion dictionaries for both fine-grained and coarse-grained sentiment analysis

Type	Opinion Dictionary	Positive	Negative
Polarity	GermanSentiStrength	69.6%	71.0%
Polarity	SentiWS	63.0%	63.0%
Strength	GermanSentiStrength	49.0%	58.2%
Strength	SentiWS	44.6%	53.2%

for German sentiment analysis: GermanPolarityClues (Waltinger, 2010), and SentiWS (Remus et al., 2010). The former includes a set of German words with binary polarity labels. Since this dictionary has no sentiment degree, it is not a good resource for our task. The latter, however, is a weighted dictionary in which each opinionated word is associated with a weight]0..1] presenting its opinion degree. We utilized this dictionary, which is a German version of SentiWordNet, in the system and presented the result in Table 9. Comparing SentiWS with our dictionary, we can see the superiority of the new dictionary provided for our study to the available opinion dictionary.

6. Conclusion

We presented a fine-grained sentiment analyzer for detecting the polarity strength of German texts using a rule-based approach. To this aim, an opinion dictionary is provided for German which contains 1,864 sentiment words as well as their strength. In addition, we provided a set of 54 booster and 17 negation words for German.

To evaluate our model, a set of 500 texts from social media about German celebrities are annotated in a fine-grained fashion. The inter-annotator agreement on the data indicates the difficulty of sentiment analysis, especially when a fine-grained annotation should be done by the system. Nevertheless, our rule-based system outperforms the other state-of-the-art approaches on both polarity level and fine-grained level.

Moreover, the introduced opinion dictionary is compared with other available German dictionaries and the results verifies the superiority of this dictionary to the other ones.

7. Acknowledgments

The author would like to thank Felix Naumann for the valuable comments and supports. The author is also very thankful to Mike Thelwall for the SentiStrength toolkit and the very interesting discussions. Additionally, Miriam Keilbach from Celebrity Performance Index Company is acknowledged for her collaboration in providing the dataset. Matthias Pohl, Erik Wendt, and Steffen Mielke are also acknowledged for annotating the dataset. Moreover, the author appreciates the work that was done by Hannes Pirker and Fabian Eckert for post processing the opinion dictionary.

8. References

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: a Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 417–422.

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS a Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting Semantic Orientations of Words Using Spin Model. In *Proceedings of the Internation Conference of the Association for Computational Linguistics (ACL)*, pages 133–140.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment Strength Detection in Short Informal Text. *American Society for Information Science and Technology*, 61(12):2544–2558.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment Strength Detection for the Social Web. *American Society for Information Science and Technology*, 63(1):163–173.

Ulli Waltinger. 2009. Polarity Reinforcement: Sentiment Polarity Identification by Means of Social Semantics. In *Proceeding of the IEEE African Communication Conference (Africon)*.

Ulli Waltinger. 2010. GermanPoralityClues: A Lexical Resource for German Sentiment Analysis. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210.

Michael Wiegand, Saeedeh Momtazi, Stefan Kazalski, Fang Xu, Grzegorz Chrupała, and Dietrich Klakow. 2008. The Alyssa System at TAC QA 2008. In *Proceeding of the Text Analysis Conference (TAC)*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the Internation Conference on Human Language Technology / Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*.